J. Dairy Sci. 99:4907–4927
http://dx.doi.org/10.3168/jds.2015-10032

Comparison of model predictions with measurements: A novel model-assessment method

N. R. St-Pierre[1]
Department of Animal Sciences, The Ohio State University, Columbus 43210

ABSTRACT

Frequently, scientific findings are aggregated using mathematical models. Because models are simplifications of the complex reality, it is necessary to assess whether they capture the relevant features of reality for a given application. An ideal assessment method should (1) account for the stochastic nature of observations and model predictions, (2) set a correct null hypothesis, (3) treat model predictions and observations interchangeably, and (4) provide quantitatively interpretable statistics relative to precision and accuracy. Current assessment methods show deficiencies in regards to at least one of these characteristics. The method being proposed is based on linear structural relationships. Unlike ordinary least-squares, where the projections from the observations to the regression line are parallel to the y-axis and inverse regression where they are parallel to the x-axis, the generalized projection regression method (GePReM) projects the observations on a regression line in a direction determined by the ratio of the precision of the observations to that of the mathematical model predictions. Estimation and testing issues arise when the model is expressed in the common slope-intercept format. A polar transformation circumvents these issues. The parameter for the angle between the regression line and the horizontal axis has symmetrical confidence intervals and is equivariant to the exchange of X and Y. The null hypothesis for the equivalence test is that the model predictions are not equivalent to the observations. Information size is calculated as the simple ratio of the variance of the true values of the observations and of the computer model predictions divided by their respective precision. This information size plays a critical role in determining the number of observations required and the size of the zone of practical tolerance for the equivalence tests. The terminology used in the comparison of measurement methods is adapted to that of model assessment based on the equivalence tests on the relative precision, regression slope, and mean bias. Two examples are presented, with complete details of the calculations required for parameter estimation, equivalence tests, and confidence intervals. The assessment method proposed is an alternative to other assessment methods available. Further research is required to establish the relative benefits and performance of this proposed method compared with others available in the literature.
Key words: model assessment, model validation, generalized projection regression method (GePReM)

INTRODUCTION

Frequently, science leads to hypotheses and theories that are best expressed using the language of mathematics. The mathematics involved can be as simple as a single function or much more complex, resulting in what are known as mathematical models. Such models can take many forms and be classified as dynamic or static, mechanistic or empirical, deterministic or stochastic (Thornley and France, 2007). Because models are abstractions and simplifications of the much more complex reality, they cannot fully characterize reality in its most intricate details. This leads to the inevitable need to assess the adequacy of a given model in representing sufficiently well the features of the real world relevant to a defined task or set of objectives. This is the essence of model assessment.

In mathematical modeling work, the model is often constructed and parameterized using domain expertise and small data sets. Eventually, external research data (i.e., data not used in model identification and parameterization) become available. These data are then used to assess the model's properties. This situation is quite different from the traditional statistical one, where the same data are used for model identification, parameterization, and model assessment.

Many methods of model assessment have been proposed and most were recently reviewed by Tedeschi (2006). In general, methods fall into one of the follow-

Received June 30, 2015.
Accepted February 21, 2016.
[1]Corresponding author: st-pierre.8@osu.edu

ing categories: linear regression (Mayer et al., 1994), including orthogonal regression (Warton et al., 2006) and modified regression (St-Pierre, 2003); analyses of deviations (Mitchell, 1997); analyses of residuals (Draper and Smith, 1988); concordance correlation coefficient (Lin, 1989); mean square error of prediction (MSEP; Bibby and Toutenburg, 1977), partitioning of MSEP into error in central tendency (i.e., mean bias), errors due to regression (i.e., linear bias), and errors due to disturbances (or random errors; Theil, 1961). All of these methods of model assessment suffer from one or more deficiencies in that they either set an incorrect model, test an incorrect hypothesis, provide metrics that are not easily interpretable, or fail to answer the right question. In addition, a useful model assessment method should provide, a priori, the characteristics of the data necessary to a useful model assessment, something akin to an a priori power determination before conducting an experiment.

The objectives of this paper are (1) to identify the most important characteristics of an ideal model assessment method, (2) to present a novel method of model assessment that meets all these characteristics, and (3) to show its application using 2 examples, the first consisting of DMI predictions in growing dairy goats (NRC, 2007), and the second dealing with predictions of microbial N flow to the duodenum of dairy cows (NRC, 2001). The new method, the generalized projection regression method (**GePReM**), will be presented without any mathematical proofs. The GePReM sets a statistical model, whereas the assessment process is for a mathematical model. The statistical model yields predictions and so does the mathematical model. To avoid confusion between the 2 models, we will refer to the mathematical model, the one being assessed, as "the computer model" in the balance of this paper, and its predictions as "the computer model predictions," although it should be clear that a mathematical model does not necessarily require a computer to yield predictions.

DESIRABLE CHARACTERISTICS OF AN IDEAL ASSESSMENT METHOD

An ideal computer model-assessment method should exhibit many desirable features (Tedeschi, 2006). Among all the desirable characteristics, arguably the most important ones can be stated as follows.

Accounting for the Stochastic Nature of Observations and Predictions

All measurements and computer models have inherent uncertainty (i.e., errors). Often the uncertainty in the predictions is not explicitly acknowledged by the model developers and is not incorporated in the computerized form of the model, but overlooking uncertainty and errors does not negate their existence.

Simple computer models can mathematically be represented by the following set of undefined functions:

$$\mathbf{Y} = \mathbf{f}(\mathbf{X}, \mathbf{B}) + \mathbf{e}, \quad [1]$$

where $\mathbf{Y}$ is a vector of n observations, $\mathbf{f}$ is a set of undefined functions, $\mathbf{X}$ is an $n \times p$ matrix of input variables, $\mathbf{B}$ is a vector of parameter estimates, and $\mathbf{e}$ is an n vector of residual errors. In this notation, stochasticity enters the computer model in many ways. First, the values of the input variables $\mathbf{X}$ are seldom known with certainty. For example, the weight of an animal when used to estimate DMI is not perfectly known. Second, the vector $\mathbf{B}$ refers to estimates of the true parameters $\boldsymbol{\beta}$, which themselves are seldom (if ever) known. By definition, the statistical estimation of parameters implies uncertainty represented by a matrix of variances and covariances of the estimated values. Third, the functional forms in $\mathbf{f}$ are rarely known with certainty. Sometimes they can be based on prevailing theories (e.g., Michaelis-Menten kinetics); many times, they are chosen among a set of candidate functions based on best-fit statistics. Hence, there is generally uncertainty regarding the specific functional forms that were chosen. Last, the residual errors cannot be ignored post-estimation. This error (uncertainty) would remain even in a perfect world, where $\mathbf{f}$, $\mathbf{X}$, and $\mathbf{B}$ would be errorless. In short, all computer model predictions are truly stochastic. Estimating prediction errors from computer models is not trivial (Marino et al., 2008). Analytical solutions are seldom available, but numerical methods such as Monte Carlo methods can generally be used quite successfully (e.g., St-Pierre and Thraen, 1999).

As for observations, their errors are generally intuitive and have been recognized in most, if not all, computer model-assessment methods.

Setting a Correct Null Hypothesis

If the comparison involves a set of parameters θ, the significance test should not be based on the conventional set of hypotheses:

$$\mathrm{H}_0\text{: } \theta = \theta_0 \text{ versus } \mathrm{H}_1\text{: } \theta \neq \theta_0. \quad [2]$$

That is, the computer model predictions should not be deemed equal to the observations unless there is enough evidence to the contrary. Instead, hypothesis tests should be set as in equivalence studies:

$$H_0: \theta \leq \theta_0 - \psi_1 \text{ or } \theta \geq \theta_0 + \psi_2 \text{ versus}$$
$$H_1: \theta_0 - \psi_1 \leq \theta \leq \theta_0 + \psi_2, \quad [3]$$

where ψ_1 and ψ_2 are parameters used to set the range of acceptable values for θ. In [3], the computer model is deemed inadequate unless it can be shown to be within a predetermined range (ψ_1 on the low side and ψ_2 on the high side) of the observations. In short, the null hypothesis should not be that the computer model and the observations are equivalent, but that they are not, with an alternate hypothesis that they are equivalent within a predetermined acceptable error. What is acceptable (i.e., the values of ψ_1 and ψ_2 in [3]) is not a statistical question but one to be settled by domain experts based on the nature of the model and its intended use.

Figure 1 illustrates the difference between the 2 sets of hypotheses for a given parameter θ. In both panels, θ_0 is the value of a parameter θ that would indicate a perfect equivalence between observations and the computer model predictions. This parameter θ is estimated using data. Figure 1a illustrates the outcome of 2 analyses using the conventional set of hypotheses. The first analysis (A) is conducted using very poor data (small number of observations, large errors, or both). Consequently, θ is estimated poorly and $\hat{\theta}_A$ has a very wide distribution. A conventional test would conclude that $\hat{\theta}_A$ is not significantly different from θ_0 because its distribution overlaps θ_0 too much. Thus, the computer model would erroneously be considered equivalent to the observations. The second analysis is conducted using extensive data with small errors. Consequently, $\hat{\theta}_B$ is very well estimated, with a very narrow confidence range. Because its distribution has a small overlap with θ_0, a conventional test on $\hat{\theta}_B$ would conclude a significant difference between $\hat{\theta}_B$ and θ_0, with the result that computer model B would not be considered equivalent to the observations. This is especially odd because computer model B produces predictions that are, on average, much closer to the observations than those of computer model A. What is even more disturbing is the realization that an analyst who wants to demonstrate the predictive quality of their model would be rewarded by using a poor data set for model assessment.

In contrast, Figure 1b illustrates an equivalence test. The perfect equivalence is still θ_0, but we add 2 boundaries, $\theta_L = \theta_0 - \psi_1$ and $\theta_U = \theta_0 + \psi_2$ to indicate a region (between lower and upper values θ_L and θ_U) for which θ would be considered close enough to θ_0 to be deemed its practical equivalent. In Figure 1b, the estimate $\hat{\theta}_A$ falls within this equivalence region but its confidence interval overlaps θ_L too much, indicating an unacceptable probability that θ is in fact smaller than θ_L. Hence, the equivalence test would indicate nonequivalence between the computer model and the observations in this case. In contrast, $\hat{\theta}_B$ also falls between θ_L and θ_U, but its confidence interval does not overlap much over the 2 boundaries. Therefore, the computer model and the observations would be considered equivalent in this case. The equivalence set of hypotheses rewards the use of good data (large number of observations and small errors) for model assessment. The boundaries θ_L and θ_U are determined by the requirements of a given application and, thus, should be set before the assessment.

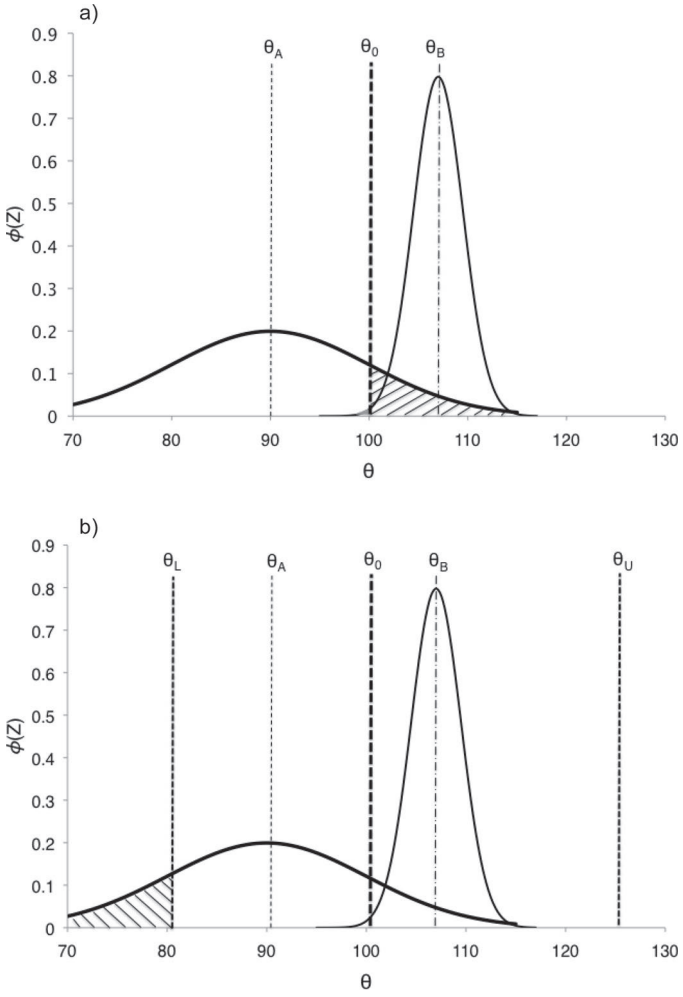

Figure 1. Comparison of a conventional difference test (a) with an equivalence test (b). In (a), θ_0 is the value of the parameter that would indicate perfect equivalence. The first estimate, θ_A, has a wide density function [ϕ(Z)] and is not statistically different from θ_0. The second estimate, θ_B, has a very narrow density function and is statistically different from θ_0. In (b), 2 boundaries (lower, L, and upper, U) of practical equivalence are added: θ_L and θ_U. The first estimate, θ_A, has a density function that considerably overlaps θ_L. Hence, θ_A does not indicate practical equivalence. The second estimate, θ_B has a density function that has very small overlaps with θ_L and θ_U. Therefore, θ_B would indicate practical equivalence.

With time, the scientific community could reach a consensus for default values to be used. At a minimum, investigators need to clearly state the boundaries used and justify their values.

Treating Model Predictions and Observations Interchangeably

Fundamentally, what we call observations and computer model predictions are, in essence, just 2 different systems of converting some inputs into outputs. The determination of the weight of an animal at a certain time will be used here as a simple example. The observation would generally be taken using a scale. This instrument converts a mechanical force (gravity) into an electronic signal using load cells (or springs and rods in old-fashioned mechanical scales). An electronic readout then converts the electronic signal into a numerical value. The scale produces an estimate of the animal's weight but with a certain error based on the precision of the scale, its calibration, and possibly small, unknown random effects, such as urination just before entering the scale. Likewise, the weight of the animal can be estimated using a model whose prediction is based on the thoracic circumference of the animal. A measuring tape is used and the linear measurement is transformed into an estimate of the weight using a predetermined linear regression that has embedded into its predictions the 4 sources of errors that were discussed previously. Both the observation and the model prediction contain errors. Both are, in fact, transformations of inputs into outputs. Deciding which one to call X and which one to call Y should be completely incidental and should have no bearing on the results of the assessment. This means that the results of an assessment should be equivariant to the exchange of X and Y. In this example, the usefulness of the measuring tape versus the scale would be determined through the assessment of their practical equivalence. That is, we would need to answer the following question: are the 2 methods of weight estimation equivalent? If not, what are their relative precision and accuracy? Perhaps surprisingly, a good measuring tape may produce more precise and accurate estimates of body weight than a bad scale.

Providing Quantitatively Interpretable Statistics on Precision and Accuracy

Not only should the method provide testable statistics related to precision and accuracy, but the statistics in question should be quantitatively interpretable. For example, the Pearson product-moment correlation r has been used as a measure of the precision of a method (Van Belle, 2002). An $r = 0.6$ is at least numerically better than an $r = 0.5$, for example, but is a measurement method with an $r = 0.6$ sufficiently precise? A scientist who is measuring the mass of different objects in kilograms would like to know the precision of the method in kilograms, not as a statistic that can take a value between −1 and 1. Likewise, the statistic used for accuracy should be expressed either in the unit of measurement (i.e., ±0.5 kg) or as a percentage (or proportion) of the measurement (i.e., ±0.5%). For a given application, a computer model with an accuracy of ±10% compared with observations might be sufficient, whereas another application might require an accuracy of ±0.1%.

GENERALIZED PROJECTION REGRESSION METHOD

The novel method, GePReM, proposed here for the assessment of computer models has its roots in error-in-variable regression methods and uses a projection that is dependent on the relative precision of the observations to the precision of the predictions from the computer model (St-Pierre, 2015a).

Establishing the Model

A statistical model where all variables are continuous and measured with errors is commonly referred as errors-in-variables regression (Casella and Berger, 1990). For the assessment method, the model is set as

$$\begin{aligned} X_i &= \xi_i + \delta_i \\ Y_i &= \eta_i + \varepsilon_i \qquad i = 1, 2, ..., n \\ \eta_i &= \alpha + \beta\xi_i, \end{aligned} \qquad [4]$$

where X_i are the computer model predictions, Y_i are the observations, ξ_i and η_i are the unobservable parameters ("true values") of X_i and Y_i, respectively, and α and β are regression parameters to be estimated. The remainders δ_i and ε_i are the errors for the computer model predictions and the observations, respectively. In this model, the 2 variables play a symmetric role. In general, the errors δ_i and ε_i are assumed to be bivariate normal; that is,

$$\begin{pmatrix} \delta_i \\ \varepsilon_i \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_\delta^2 & 0 \\ 0 & \sigma_\varepsilon^2 \end{bmatrix}\right), \qquad [5]$$

where σ_δ^2 is the precision parameter (i.e., variance) of the computer model predictions, and σ_ε^2 is the precision parameter (i.e., variance) of the observations. It must be noted that in some fields of study, precision is ex-

pressed as the reciprocal of the variance. Such is not the case here to remain consistent with the relevant literature (e.g., Solari, 1969; Anderson 1976, 1984; Casella and Berger, 1990). As a linear model, [4] is commonly referred to as a linear statistical relationship (Anderson, 1984). When ξ_i and η_i are considered fixed, the model is called a linear functional relationship, whereas when they are considered random, it is called a linear structural relationship (Fuller, 1987).

Parameter Estimates

Maximum likelihood (**ML**) estimates of all parameters in [4] do not exist unless we assume prior knowledge of the variances σ_δ^2 and σ_ε^2 (Kendall and Stuart, 1979). The error variance for the measurements $\left(\sigma_\varepsilon^2\right)$ can be estimated relatively easily from repeated measurements on the same unit. For example, if we use the average daily DMI obtained over a week as a measurement, then the variance of the 7 daily measurements divided by 7 could serve as a good estimate of σ_ε^2 if the process (i.e., DMI) is relatively stationary through the 7 d of measurements. An estimate of σ_δ^2 is explicit when the computer model is set as a stochastic model, which unfortunately is not very often. However, it is relatively easy to use Monte Carlo methods to account for the variances and covariances of **X**, **B**, and **e** implied in [1] to generate a distribution for a prediction by the computer model, hence allowing a good estimate of σ_δ^2. Furthermore, it has been shown that it is sufficient to know the ratio of σ_ε^2 to σ_δ^2 for ML estimates of [4] to exist (Solari, 1969; Kendall and Stuart, 1979; Willassen, 1979). Thus, we only need to know the relative precision of the measurements compared with computer model predictions to obtain ML estimates of α and β in [4]. This ratio, λ, where

$$\lambda = \frac{\sigma_\varepsilon^2}{\sigma_\delta^2}, \qquad [6]$$

is called the *precision ratio*. It plays an important role in the estimation of the parameters in [4], and in comparing the relative precision of the measurements compared with the predictions by the computer model. For example, $\lambda = 2$ indicates that the computer model predictions are twice as precise as the measurements, whereas $\lambda = 0.5$ would indicate that the measurements are twice as precise as the computer model predictions. Hypotheses on λ can easily be tested because $\hat{\lambda}/\lambda$ has an F ratio (Tan and Iglewicz, 1999). Given λ, the ML estimate of β in [4] is (Kendall and Stuart, 1979)

$$\hat{\beta} = \frac{S_{yy} - \lambda S_{xx} + \sqrt{\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S_{xy}^2}}{2S_{xy}}, \qquad [7]$$

where $S_{xx} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$, $S_{yy} = \sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2$, and $S_{xy} = \sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)$, and $\bar{X}$ and $\bar{Y}$ are the means of X and Y, respectively.

The estimate of β in [7] has also been known as a Deming regression or total least-squares and is a compromise between the slope estimates of the simple regression of Y on X and that of X on Y (Tan and Iglewicz, 1999). When $\lambda \to \infty$; that is, when the computer model prediction errors approach 0, the projection line becomes vertical and $\hat{\beta} \to S_{xy}/S_{xx}$, which is the solution to the simple ordinary least-squares regression (**OLS**) of Y on X. Likewise, when $\lambda \to 0$; that is, the observation errors approach 0, then the projection line approaches horizontal and $\hat{\beta} \to S_{yy}/S_{xy}$, which is the solution for the regression line of X on Y known as inverse regression (**IR**). Thus, [7] provides the solution to a generalized projection line determined by the ratio of precisions (i.e., λ). The solution to [7] is bounded by the solutions to the simple regression of Y on X and the inverse regression of X on Y. It should also be noted that when $\lambda = 1$, the ML estimate of β simplifies to

$$\hat{\beta} = \frac{S_{yy} - S_{xx} + \sqrt{\left(S_{yy} - S_{xx}\right)^2 + 4S_{xy}^2}}{2S_{xy}}, \qquad [8]$$

which is simply the solution found by orthogonal least squares regression (**OR**), so named because the projection line (i.e., the line linking an observation to its prediction) is orthogonal (perpendicular) to the fitted line (Madansky, 1959; Casella and Berger, 1990; Carroll and Ruppert, 1996). Thus, [7] reduces to an OR solution when the observations and the computer model predictions have the same precision.

The ML estimate of α in [4] is simply (Kendall and Stuart, 1979):

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}. \qquad [9]$$

Because the GePReM regression goes through $\left(\bar{X}, \bar{Y}\right)$, the mean bias in model assessment $\left(\bar{B}\right)$ is simply estimated as $\hat{\bar{B}}$ (Altman and Bland, 1983):

$$\hat{\bar{B}} = \bar{X} - \bar{Y}. \qquad [10]$$

Confidence Intervals

Confidence Intervals for β when λ = 1. Confidence intervals (**CI**) for the $\hat{\beta}$ estimated by [7] are not symmetric about $\hat{\beta}$ (Tan and Iglewicz, 1999) and can have infinite expected length when derived without further restrictions in the scale of $\hat{\beta}$ (Gleser and Huang, 1987). Fortunately, Creasy (1956) proposed an alternate format for model [4] that will allow us to circumvent this problem. A straight line can be expressed in the common slope-intercept format, but it can also be expressed in polar form as

$$\eta_i \cos\theta + \xi_i \sin\theta = \tau \qquad \left(-\pi / 2 < \theta < \pi / 2\right), \qquad [11]$$

where θ is the angle (in radians) between the line of regression and the horizontal axis. It is then easy to establish a set of connections between the parameters in [7] and those in [11]:

$$\begin{aligned} \beta &= \tan\theta \qquad & \theta &= \arctan\beta \\ \alpha &= \frac{\tau}{\cos\theta} \qquad & \tau &= \operatorname{sgn}(\beta)\frac{\alpha}{\sqrt{1+\beta^2}} \end{aligned} \qquad [12]$$

The transformation to a polar form allows the derivation of an exact CI for θ. Under certain conditions to be explained shortly, the CI for β can be expressed by taking the tangent of the lower and upper confidence limits of θ (i.e., applying [12]). When $\lambda = 1$ (i.e., observations and computer model predictions have the same precision), the $100(1 - \gamma)\%$ CI for θ is (Creasy, 1956)

$$\hat{\theta} - \phi_{\gamma/2} < \theta < \hat{\theta} + \phi_{\gamma/2}, \qquad [13]$$

with

$$\phi_{\gamma/2} = \frac{1}{2}\arcsin\left(t_{n-2,\gamma/2} \times \frac{2}{\sqrt{n-2}} \times \sqrt{\frac{S_{yy}S_{xx} - S_{xy}^2}{\left(S_{yy} - S_{xx}\right)^2 + 4S_{xy}^2}}\right), \qquad [14]$$

where $t_{n-2,\gamma/2}$ is the $100(1 - \gamma/2)\%$ percentile of the t distribution with $n - 2$ df. The CI in [13] is equivariant to exchange of X and Y and its boundaries are symmetric about $\hat{\theta}$ (Creasy, 1956).

CI for β when λ ≠ 1. The CI for θ in [13] are correct only for the restricted case when $\lambda = 1$. For the more general case, where the precision ratio λ is not equal to 1, we first need to define the relative sensitivity of Y with respect to X (β_S; Mandel, 1978) as

$$\beta_S = \frac{\beta}{\sqrt{\lambda}}. \qquad [15]$$

Later, I will expand on the interpretation of β_S in model assessment. For now, [15] can be used to simplify the model using the following transformation (Tan and Iglewicz, 1999):

$$X' = \sqrt{\lambda}X. \qquad [16]$$

In [16] X' are the values of the model predictions in a transformed scale defined by the square root of the precision ratio. It is easy to demonstrate that the precisions of X' and Y are the same after this transformation. Put differently, $\sigma_{\delta'}^2 = \sigma_{\varepsilon}^2$ and $\lambda' = 1$. Then we can use [7] with X' replacing X, and obtain a new slope $\beta_S = \beta/\sqrt{\lambda}$ in the scale of the transformed variables. We can also define $\theta_s = \arctan(\beta_S)$ so that

$$\hat{\theta}_s = \arctan\left(\hat{\beta}_S\right) = \arctan\left(\frac{\hat{\beta}}{\sqrt{\lambda}}\right). \qquad [17]$$

Using [17] into [13], we get the following $100(1 - \gamma)\%$ confidence interval for θ_s:

$$\hat{\theta}_s - \phi_{\gamma/2}(\lambda) < \theta_s < \hat{\theta}_s + \phi_{\gamma/2}(\lambda), \qquad [18]$$

with

$$\phi_{\gamma/2}(\lambda) = \frac{1}{2}\arcsin\left(t_{n-2,\gamma/2} \times \frac{2}{\sqrt{n-2}} \times \sqrt{\frac{\lambda\left(S_{yy}S_{xx} - S_{xy}^2\right)}{\left(S_{yy} - \lambda S_{xx}\right)^2 + 4\lambda S_{xy}^2}}\right). \qquad [19]$$

The inequality in [18] provides the confidence interval for θ_s. From there, it is easy to convert the lower and upper bounds for this interval into an interval for β. Because $\beta = \sqrt{\lambda}\beta_S = \sqrt{\lambda}\tan(\theta_s)$, the $100(1 - \gamma)\%$ CI for β is

$$\sqrt{\lambda}\tan\left[\hat{\theta}_s - \phi_{\gamma/2}(\lambda)\right] < \beta < \sqrt{\lambda}\tan\left[\hat{\theta}_s + \phi_{\gamma/2}(\lambda)\right]. \qquad [20]$$

Both the CI for θ_s given by [18] and the CI for β by [20] are equivariant to exchange of X and Y.

CI for Mean Bias. The standard error of mean bias $\bar{B}$ is calculated as:

$$SE_{\bar{B}} = \sqrt{\frac{S_{XX} + S_{YY} - 2S_{XY}}{n(n-1)}}. \qquad [21]$$

The $100(1-\gamma)\%$ CI for $\bar{B}$ is simply

$$\hat{\bar{B}} - t_{n-2,\,\gamma/2}\; SE_{\hat{B}} < \bar{B} < \hat{\bar{B}} + t_{n-2,\,\gamma/2}\; SE_{\hat{B}}, \qquad [22]$$

where $t_{n-2,\gamma/2}$ is the $100(1-\gamma/2)\%$ percentile of the t distribution with $n-2$ df.

Reasonable Lengths of CI on β

Unfortunately, the CI provided by [18] can have infinite expected length (Gleser and Huang, 1987). The source of the problem can be identified. The CI in [13] and [18] are determined by $\phi_{\gamma/2}$, which itself involves the arcsine function (i.e., [19]). This function is defined only when the expression inside the parentheses takes a value $\leq |1|$. At its limit, $\arcsin(1) = \pi/2$, resulting in the following CI for β in equation [20]: $-\infty < \beta < \infty$. Clearly, this CI is not informative. Therefore, we need to define the conditions for which the expression within the arcsin function takes a value sufficiently smaller than |1| to result in meaningful CI.

To do this, we must first introduce a new statistic, the *information size*, denoted by κ^2. For structural relationships,

$$\kappa^2 = \frac{\sigma_\xi^2}{\sigma_\delta^2} + \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}, \qquad [23]$$

where σ_ξ^2 is the variance of the true values from the computer model $\left(\text{i.e., } \xi_i\right)$, and σ_η^2 is the variance of the true values for the observations $\left(\text{i.e., } \eta_i\right)$. In [23], we can see that κ^2 is simply the sum of 2 ratios of the variance (i.e., spread) of the true values for the observations and the computer model predictions divided by their respective precision. This information size plays a critical role in obtaining meaningful CI for β because the problem of a wide CI occurs only when κ^2 is close to zero (Tan and Iglewicz, 1999). Therefore, we need a lower bound on κ^2 that will prevent getting uninformative CI. Tan and Iglewicz (1999) derived the conditions for which κ^2 is sufficiently large to obtain reasonable CI with high probability. When $\lambda = 1$ in a structural relationship, if κ^2 satisfies the following condition:

$$\kappa^2 > \frac{\left(2t_{n-2,\gamma/2} + \sqrt{n-2}\sin\omega_\theta\right)}{\left(n-1\right) f_{n-1,n-2,1-\nu}\sin^2\omega_\theta} - 1, \qquad [24]$$

where $f_{n-1,n-2,1-\nu}$ is the $100\nu\%$ percentile of an F distribution with $n-1$ and $n-2$ df, then with probability at least $1-\nu$, the length of the $100(1-\gamma)$ confidence for θ is less than ω_θ (with the condition that $0 < \omega_\theta < \pi/2$). The parameter ω_θ in [24] is the width of the CI of θ. We can think of an acceptable width as 2 times the difference from $\theta = \pi/4 \approx 0.7854$ (i.e., exact equivalence) that would be deemed satisfactory. For example, we could be satisfied if the 95% CI of θ had a width equal to 95% of the value of θ if we had perfect equivalence. In which case, we would have $\omega_\theta = \pi/4 - (0.95 \times \pi/4) \approx 0.0393$. Table 1 provides the lower bounds for κ^2 for various ω_θ when $\lambda = 1$. Recall that the confidence limits are symmetric to θ but not to β. Most people have a better intuitive sense of what β means (i.e., should be equal to 1 if the observations and the computer model predictions are equivalent) than of θ (i.e., $\theta = \pi/4 \approx 0.7854$ under equivalence). Therefore, Table 1 reports the lower bounds on κ^2 for various widths of CI on the β scale (which we denote as ω_β) with ω_β set at values so that the widths of the 95% CI on β are equal to 0.2, 0.1, 0.05, and 0.01 with a 95% probability. For example, suppose that we have 100 observations available (i.e., $n = 100$), and we want to ensure that we will have a 95% CI on β with a width $\omega_\beta \leq 0.2$. Table 1 indicates that for $\omega_\beta = 0.2$, a $\kappa^2 \geq 144$ would ensure that the 95% CI on β would have a width ≤0.2 with a 95% probability. Of course, Table 1 can also be used to estimate the number of observations required for a given κ^2 and a desired width ω_β.

The inequality in [24] is valid only when $\lambda = 1$ (i.e., the observations and the computer model predictions have equal precision). Using the transformation in [16] makes the inequality applicable in the scale of θ_S when $\lambda \neq 1$, but then we need to determine the correspondence of the interval length in θ_S, which we denote as ω_S, to that of ω_θ and ω_β. The exact translation of the interval length in θ (i.e., ω_θ) into length on θ_S (i.e., ω_S) depends on the value of θ (Tan and Iglewicz, 1999). However, if the computer model is to be useful, then θ should be around $\pi/4$ (i.e., $\beta \approx 1$). Conditional to the value of $\theta = \pi/4$, the correspondence of the interval length is

$$\omega_S = \frac{2\sqrt{\lambda}}{\lambda+1} \times \omega_\theta. \qquad [25]$$

Table 1 provides the lower bounds on the information size κ^2 for $\lambda = 2$, 4, and 8 using equation [24] with ω_S replacing ω_θ. Going back to our prior example, if a model is to be assessed where 100 observations are available and we want a 95% CI on β to have a width ≤0.2 with a 95% probability, then the information size κ^2 needs to be at least 144, 160, 216, and 339, when $\lambda = 1$, 2, 4, and 8, respectively. The lower bound on information size goes up as λ increases. It is evident that

Table 1. Low bounds on information size (κ^2) at different sample sizes (n) and precision ratios (λ) for the 95% CI of θ_S, θ, and β[1] of having at least a 95% probability within the given widths ω_S, ω_θ, and ω_β[2]

Precision ratio	ω_S	ω_θ	ω_β	Sample size, n				
				25	50	100	250	500
$\lambda = 1$								
	0.100	0.100	0.20	847	326	144	55	28
	0.050	0.050	0.10	3,218	1,214	520	190	94
	0.025	0.025	0.05	12,529	4,673	1,971	701	338
	0.005	0.005	0.01	307,149	113,450	47,240	16,392	7,699
$\lambda = 2$								
	0.0943	0.100	0.20	948	365	160	61	31
	0.0471	0.050	0.10	3,611	1,360	582	212	105
	0.0236	0.025	0.05	14,076	5,246	2,211	785	378
	0.0047	0.005	0.01	345,433	127,573	53,111	18,422	8,649
$\lambda = 4$								
	0.0800	0.100	0.20	1,299	496	216	81	41
	0.0400	0.050	0.10	4,978	1,869	796	288	141
	0.0200	0.025	0.05	19,481	7,248	3,048	1,077	516
	0.0040	0.005	0.01	479,426	176,996	73,651	25,651	11,971
$\lambda = 8$								
	0.0629	0.100	0.20	2,071	786	339	125	63
	0.0314	0.050	0.10	8,001	2,993	1,267	454	221
	0.0157	0.025	0.05	31,425	11,668	4,893	1,721	820
	0.0031	0.005	0.01	776,007	286,368	119,094	41,226	19,313

[1]θ_S is the slope in polar coordinates in the scale of the transformed variables, θ is the slope in polar coordinates in the untransformed scale, and β is the slope in the common slope-intercept format in the untransformed scale.

[2]Lower bounds on κ^2 in this table are calculated using equation [24] in the text using ω_θ for $\lambda = 1$, and ω_S when $\lambda \neq 1$.

the precision ratio λ plays an important role in the type and size of data required to conduct an assessment. Before attempting a model assessment, a scientist could easily determine whether the data available are both sufficient in number (i.e., n) and information size (i.e., κ^2) to yield an estimate of β with sufficient precision to be meaningful. If the data are insufficient, the scientist can (1) seek additional data (i.e., increase n); (2) increase the range of the observations and computer model predictions from new data (i.e., increase κ^2); (3) increase the precision of either (or both) the observations or the computer model predictions; that is, increasing $\sigma_\xi^2/\sigma_\delta^2$ or $\sigma_\eta^2/\sigma_\varepsilon^2$ in [23], resulting in a greater κ^2; or (4) select a greater value for $\gamma/2$ (i.e., use a 90% CI as opposed to a 95% CI).

Equivalence Tests

As pointed out previously, it is important to set the correct hypotheses for conducting an equivalence test (i.e., use the hypotheses set in [3] as opposed to the conventional set in [2]). Using the intersection union testing approach, Berger and Hsu (1996) detailed the method to be used for constructing correct equivalence tests.

Equivalence on Slope when $\lambda = 1$. When $\lambda = 1$ and $\psi_1 = \psi_2 = \psi$ in [3], the resulting test at a γ level has the following rejection region:

$$\left|\hat{\theta} - \theta_0\right| < \psi - \phi_\gamma. \quad [26]$$

In [26], the test on the slope is conducted on θ, which is the polar coordinate for expressing the slope. This is because the CI on the slope is symmetric to $\hat{\theta}$ but not to $\hat{\beta}$. The value of ψ is chosen to represent a region of acceptability for $\hat{\theta}$. This region can be translated in the slope-intercept format (i.e., for $\hat{\beta}$) using [12] if different tolerances on $\hat{\beta}$ are desired.

Equivalence Test on the Slope when $\lambda \neq 1$. The procedure that we have followed throughout makes the generalization when $\lambda \neq 1$ not overly problematic. The inequality in [26] applies when $\lambda = 1$. When $\lambda \neq 1$, we use the scale transformation in [16] to make $\lambda = 1$ in the transformed scale for X and then apply the procedure derived for $\lambda = 1$ to the transformed data. Simply put, [26] becomes

$$\left|\hat{\theta}_S - \theta_{S_0}\right| < \psi_S - \phi_\gamma(\lambda). \quad [27]$$

where θ_{S_0} is the value of θ_S that equals perfect equivalence. The computer model predictions are perfect equivalent to the observations if $\beta = 1$. Because $\theta_S = \arctan\left(\hat{\beta}/\sqrt{\lambda}\right)$, [27] can be rewritten as

$$\left|\hat{\theta}_S - \arctan\left(\frac{1}{\sqrt{\lambda}}\right)\right| < \psi_S - \phi_\gamma(\lambda). \quad [28]$$

Table 2. Values of ψ_S to be used for constructing confidence intervals of θ_S and equivalence tests to control the lower limit of practical equivalence (β_L) at various precision ratios (λ)[1]

	β_L				
λ	0.8	0.9	0.95	0.975	0.99
1.0	0.110657	0.052583	0.025635	0.012658	0.005025
1.5	0.106100	0.050987	0.024990	0.012371	0.004919
2.0	0.100674	0.048727	0.023965	0.011884	0.004730
2.5	0.095535	0.046471	0.022911	0.011375	0.004530
3.0	0.090909	0.044382	0.021921	0.010893	0.004341
3.5	0.086797	0.042493	0.021017	0.010451	0.004167
4.0	0.083141	0.040794	0.020199	0.010050	0.004008
4.5	0.079879	0.039264	0.019459	0.009686	0.003864
5.0	0.076954	0.037881	0.018788	0.009356	0.003733
5.5	0.074314	0.036627	0.018178	0.009055	0.003614
6.0	0.071920	0.035485	0.017620	0.008779	0.003504
6.5	0.069736	0.034439	0.017109	0.008527	0.003404
7.0	0.067736	0.033478	0.016638	0.008294	0.003311
7.5	0.065895	0.032591	0.016203	0.008078	0.003226
8.0	0.064194	0.031769	0.015800	0.007878	0.003146
8.5	0.062616	0.031006	0.015425	0.007692	0.003072
9.0	0.061148	0.030294	0.015074	0.007519	0.003003
9.5	0.059777	0.029628	0.014746	0.007356	0.002938
10.0	0.058494	0.029004	0.014439	0.007203	0.002877

[1]ψ_S defines the region of acceptability for the slope θ_S, which is expressed in polar coordinates in the scale of the transformed variables.

In [28], ψ_S is chosen in the transformed scale (i.e., for θ_S); thus it is in polar coordinates. Its correspondence in the slope-intercept form is dependent on λ. Table 2 provides the correct ψ_S to be used to correspond to a desired tolerance on the lower bound for the slope β, depending on the precision ratio. Table 3 provides the resulting upper bounds on the CI for β. For example, a $\psi_S = 0.024169$ would be used in [28] to produce an equivalence test of the form $0.95 < \beta < 1.052$ when the precision ratio $\lambda = 2$. The CI is symmetric to θ_S but not to β. The values in Table 2 are provided for round values on the lower bound for β. Consequently, the upper bounds on β (i.e., β_U) are found in Table 3. The asymmetry of the confidence bounds on β gets smaller, the tighter the desired interval (Table 3). The exact value of ψ_S to be used for a given λ and a desired lower bound for β (i.e., β_L) can be calculated using the following equation:

$$\psi_S = \arctan\left(\frac{\sqrt{\lambda} - \beta_L\sqrt{\lambda}}{\beta_L + \lambda}\right). \quad [29]$$

The exact value for B_U is then

$$\beta_U = \frac{\sqrt{\lambda} + \lambda\tan(\psi_S)}{\sqrt{\lambda} - \tan(\psi_S)}. \quad [30]$$

Equivalence Test on the Mean Bias. The equivalence test on $\overline{B}$ at a γ level has the following rejection region:

$$\left|\hat{\overline{B}}\right| < \psi_{\overline{B}} - \phi_{\gamma/2,\overline{B}}, \quad [31]$$

with

$$\phi_{\gamma/2,\overline{B}} = t_{n-2,\gamma/2}\ SE_{\overline{B}}, \quad [32]$$

and $\psi_{\overline{B}}$ is the acceptable upper limit tolerance on the absolute value of the overall bias.

Geometric Interpretation

In OLS, parameter estimates are obtained by minimizing the sum of the squared residuals calculated as the observed values minus their vertical projections on the regression line (Figure 2a). In contrast, IR minimizes the sum of the squared residuals calculated as the observed values minus their horizontal projections on the regression line (Figure 2b). Orthogonal regression minimizes the sum of the squared residuals calculated as the observations minus their perpendicular (i.e., orthogonal) projection on the regression line (Figure 2c). The lines fitted by OR always lie somewhere between the lines fitted by OLS and IR. With GePReM, X_1 is first transformed to $X_1' = \sqrt{\lambda}\ X_1$ so that the precision ratio (λ) between X_2 and X_1' is equal to 1, and OR is then conducted by regressing X_2 on X_1'. An equivalent way of picturing GePReM is to think that the direction of the projection is determined by λ. When λ is very large, the projections approach those of OLS and the GePReM solution is close to that of OLS. When λ is

Table 3. Upper limits of practical equivalence (β_U) for various lower limits of practical equivalence (θ_L) resulting from the ψ_S values listed in Table 2[1]

Precision ratio (λ)	β_L				
	0.8	0.9	0.95	0.975	0.99
1.0	1.250	1.111	1.053	1.026	1.010
1.5	1.238	1.109	1.052	1.026	1.101
2.0	1.231	1.107	1.052	1.025	1.010
2.5	1.226	1.106	1.051	1.025	1.010
3.0	1.222	1.105	1.051	1.025	1.010
3.5	1.220	1.105	1.051	1.025	1.010
4.0	1.217	1.104	1.051	1.025	1.010
4.5	1.216	1.104	1.051	1.025	1.010
5.0	1.214	1.103	1.051	1.025	1.010
5.5	1.213	1.103	1.051	1.025	1.010
6.0	1.212	1.103	1.051	1.025	1.010
6.5	1.211	1.103	1.051	1.025	1.010
7.0	1.211	1.103	1.051	1.025	1.010
7.5	1.210	1.102	1.051	1.025	1.010
8.0	1.209	1.102	1.051	1.025	1.010
8.5	1.209	1.102	1.051	1.025	1.010
9.0	1.208	1.102	1.051	1.025	1.010
9.5	1.208	1.102	1.050	1.025	1.010
10.0	1.208	1.102	1.050	1.025	1.010

[1]ψ_S is the parameter used to set the range of acceptable values for θ_S, the slope expressed in polar coordinates in the scale of the transformed variables.

very small (i.e., approaches zero), the projections approach those of IR and the GePReM solution is close to that of IR. When $\lambda = 1$, the GePReM solution is that of OR.

Residuals are calculated differently in GePReM than they would be with OLS. The X_i and Y_i (i.e., the data) in [4] contain errors in both variables. The residuals are no longer calculated as $Y_i - (\alpha + \beta X_i)$ as they would be in OLS, but are calculated as $\eta_i - (\alpha + \beta\xi_i)$; that is, using the true but unknown values of the observations and computer model predictions. A generalized projection from a data point (X_1, Y_1) to a GePReM regression line is shown in Figure 3. The residual in this instance is the Euclidean distance between the data point (X_1, Y_1) and its projection $\left(X_1^*, Y_1^*\right)$. The ML estimate of the true value of X_i for the ith data point is

$$\hat{X}_i^* = X_i + \left(\frac{\hat{\beta}}{\hat{\beta}^2 + \lambda}\right) \times \left(Y_i - \hat{\alpha} - \hat{\beta}\ X_i\right), \qquad [33]$$

and the ML estimate of the true value of Y_i is then simply

$$\hat{Y}_i^* = \hat{\alpha} + \hat{\beta}\hat{X}_i^*. \qquad [34]$$

The residual for the ith data, r_i, is then calculated as the signed distance between the data point and its projection:

$$r_i = \text{sgn}\left(Y_i - Y_i^*\right) \times \sqrt{\left(Y_i^* - Y_i\right)^2 + \left(X_i^* - Y_i\right)^2}. \qquad [35]$$

ESTIMATION OF VARIANCES

The GePReM method requires a priori knowledge of 4 variances: σ_η^2, σ_ε^2, σ_ξ^2, and σ_δ^2. The first variance, σ_η^2, is the variance of the true values of the observations. Of course, we do not know the true values η_i in [4] and, hence, their variance, but we do know the variance of Y_i in [4] (i.e., the simple variance of the "apparent" observations). Because η_i and ε_i in [4] are generally independent [i.e., the precision of the measurements is unrelated to the variance (range) of the true values], we have

$$\sigma_Y^2 = \sigma_\eta^2 + \sigma_\varepsilon^2. \qquad [36]$$

Getting an estimate of σ_ε^2 is generally not a major hurdle because it represents the precision of the measurements, something that is generally known. Therefore, the variance of the true values of the observations is simply calculated as

$$\sigma_\eta^2 = \sigma_Y^2 - \sigma_\varepsilon^2. \qquad [37]$$

In the event that the variance of Y is not known, as when using [24] and [25] before doing a model assessment to determine whether the data likely available will bring sufficient power to the test statistics, then an approximate a priori estimate of σ_Y^2 is obtained from the expected range of the observations (Hozo et al., 2005):

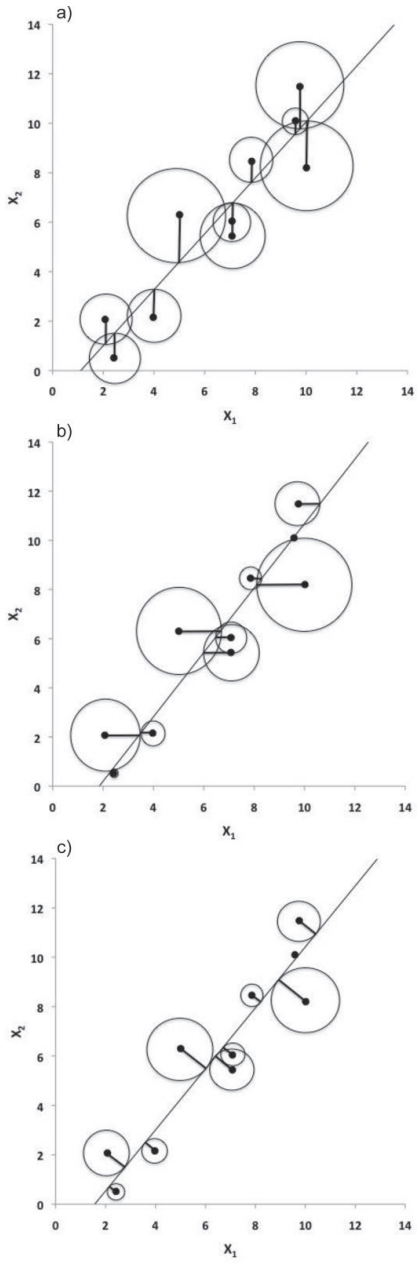

Figure 2. Geometric interpretation of standard ordinary least-squares regression (OLS), inverse regression (IR), and orthogonal regression (OR) using simulated data. Panel 2a depicts OLS, a method that minimizes the sum of squares of the vertical distances between observations and their predictions. Because the squared residuals are proportional to the surface areas of the depicted circles, the fitted line minimizes the total area of these circles. Panel 2b shows the approach used by IR. Here, the minimization is of the sum of squared of the horizontal distances between observations and predictions. In this instance, it is equivalent to minimizing the surface areas of the depicted circles, which are not equal to those of panel 2a. In panel 2c, the residuals for OR are perpendicular to the fitted line with lengths equal to the radii of the circles. Orthogonal regression minimizes the surface areas of these circles. The solution obtained by the generalized projection regression method (GePReM) is equal to OR when the precision of the model predictions is equal to that of the measurements. When the precisions differ, the GePReM solution moves toward OLS or IR depending on the precision ratio λ.

$$\sigma_Y^2 = range \div 4 \text{ if } 25 < n < 70,$$

$$\sigma_Y^2 = range \div 6 \text{ if } n > 70. \quad [38]$$

If the true values of the computer model predictions are to be close to the true values of the observations, then $\sigma_\xi^2 \approx \sigma_\eta^2$. The parameter that will likely be most difficult to obtain is the precision of the model predictions $\left(\sigma_\delta^2\right)$ because so many models are being built and reported as deterministic as opposed to stochastic models. Monte Carlo estimation methods are generally easily implemented when we have knowledge of the variances and covariances of the parameter estimates used by the computer models (i.e., the **B** in [1]) as well as for its inputs (i.e, the **X** in [1]). The uncertainty regarding variance and covariances estimates can also be factored in using Markov chain Monte Carlo as implemented, for example, in OpenBUGS (Lunn et al., 2009). Many times, educated guesses will be sufficient, but we can hope that future model developers will become more aware of the need to assess the precision of their models.

QUALIFYING EQUIVALENCY AND BIASES

The 2 fundamental performance parameters of any measurement method are accuracy and precision. I argue that the assessment of a computer model is best viewed as the comparison of 2 methods of measurements, albeit one being more direct than the other. Hence, the terminology used in the comparison of measurement

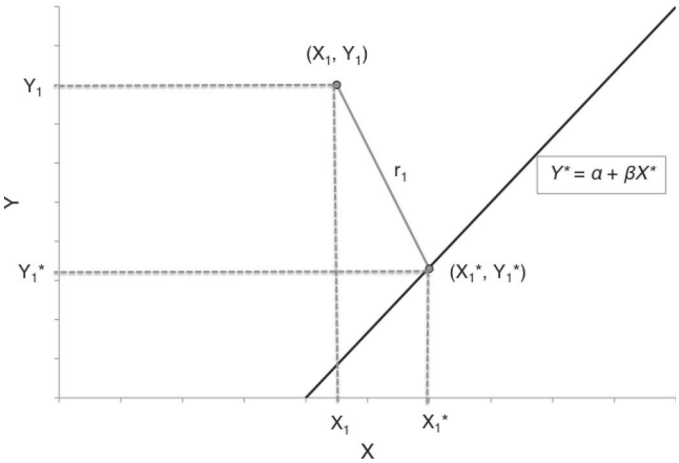

Figure 3. Graphical representation of the generalized projection and the calculation of the residual value. The point (X_1, Y_1) is the observed value. Its projection on the generalized projection regression method (GePReM) regression line is in a direction determined by the precision ratio λ. The residual r_1 is the signed distance between (X_1, Y_1) and (X_1*, Y_1*).

methods will be briefly stated. This terminology will be adapted to model assessment with the introduction of tolerances on parameter estimates, where such tolerances are deemed acceptable in practice.

Individual Equivalent

Two measurement methods are called *individual equivalent* if they have the same accuracy and precision. That is, when $\lambda = 1$, $\beta = 1$, and $\alpha = 0$. Thus, if a model is deemed individual equivalent to observations, its predictions can be used in replacement of direct observations without loss of accuracy and precision. Recall that $\beta_S = \beta/\sqrt{\lambda}$. Thus, a test on $\beta_S = 1$ would be an omnibus test on whether the computer model is an individual equivalent (a separate test on $\alpha = 0$ would still be required, but this test is orthogonal to the test on β_S and is quite trivial).

Average Equivalent

Two measurement methods are called *average equivalent* if they have the same accuracy profile but not the same precision. That is, when $\alpha = 0$, $\beta = 1$, and $\lambda \neq 1$.

Unless the computer model is an algebraic transformation of the observations, pure individual equivalent and average equivalent computer models probably do not exist. This is because a computer model always simplifies the complex reality. The relevant question, however, is whether the observations and the computer models are close enough to be equivalent in practice—what will be called practical equivalent.

Practical Individual Equivalent

The computer model will be called practical individual equivalent to observations if $\lambda = 1$ within a tolerance ψ_λ, $\beta = 1$ or equivalently $\theta = \pi/4$ with a tolerance ψ_θ, and $\overline{B} = 0$ with a tolerance $\psi_{\overline{B}}$. The predictions from a practical individual equivalent computer model can be substituted to measurements without any practical loss of precision and accuracy.

Practical Average Equivalent

The computer model will be called practical average equivalent to observations if $\lambda \neq 1$ within a tolerance ψ_λ, $\beta = 1$, which is formally stated as $\theta_S = \arctan\left(\frac{1}{\sqrt{\lambda}}\right)$, with a tolerance ψ_{θ_S}, and $\overline{B} = 0$, with a tolerance $\psi_{\overline{B}}$. In the polar coordinates, the practical equivalence for the slope is

$$\arctan\frac{1}{\sqrt{\lambda}} - \psi_S < \theta_S < \arctan\frac{1}{\sqrt{\lambda}} + \psi_S. \qquad [39]$$

Because $\beta = \sqrt{\lambda}\tan\theta_S$, the practical equivalence for the slope in the Cartesian coordinates (i.e., β) is

$$\frac{\sqrt{\lambda} - \lambda\tan\psi_S}{\sqrt{\lambda} + \tan\psi_S} < \beta < \frac{\sqrt{\lambda} + \lambda\tan\psi_S}{\sqrt{\lambda} - \tan\psi_S}. \qquad [40]$$

In short, practical average equivalence requires that the slope $\beta = 1$ within a practical tolerance and that the mean bias $\overline{B} = 0$ also within a practical tolerance. The predictions from a practical average equivalent model have a precision that is not practically the same as that of the observations, but they have an accuracy that is practically the same.

Practical Shift Equivalent

The computer model will be called practical shift equivalent to observations if $\lambda \neq 1$ within a tolerance ψ_λ, $\beta \neq 1$, which is formally stated as $\theta_S \neq \arctan\left(\frac{1}{\sqrt{\lambda}}\right)$ within a tolerance ψ_{θ_S}, and $\overline{B} = 0$, with a tolerance $\psi_{\overline{B}}$. The predictions from a practical shift equivalent model have a precision that is different than that of the observations. The predictions are also reasonably close to the observations on an average (i.e., no overall bias), but the slope β is sufficiently different from 1 to indicate that computer model predictions and observations differ more than what is practically acceptable in a portion of the prediction space.

Practical Drift Equivalent

The computer model will be called practical drift equivalent to observations if $\lambda \neq 1$ within a tolerance ψ_λ, $\beta = 1$, which is formally stated as $\theta_S = \arctan\left(\frac{1}{\sqrt{\lambda}}\right)$ within a tolerance ψ_{θ_S}, and $\overline{B} \neq 0$ within a tolerance $\psi_{\overline{B}}$. On average, the computer model predictions are not equivalent to the observations, but the differences between the computer model predictions and the observations are not dependent on the values of the prediction (i.e., no linear bias). This would be the situation, for example, if we were comparing weights measured on a scale that had not been zeroed with those taken on a properly zeroed scale.

EXAMPLES

Example 1: Prediction of DMI in Growing Goats

The observations come from 6 studies conducted at the Universidade de Estadual Paulista (UNESP) in Jaboticabal, Brazil (Teixeira et al., 2011). The complete data set is provided in Supplemental Table S1 (http://dx.doi.org/10.3168/jds.2015-10032). In total, the average daily DMI of 67 growing goats with a mean BW of 19.5 kg (range of 9.9 to 28.5 kg) and fed ad libitum was measured over an average of 76 d of growth. The model to be assessed is the digestibility-adjusted equation from NRC (2007), as stated by Teixeira et al. (2011):

$$\text{DMI} = (76.7\ \text{BW}^{0.75}) \times (-0.666 + 1.333\text{ME} - 0.0266\text{ME}^2). \quad [41]$$

The mean of the observed DMI was 795.8 g/d (range of 135 to 1,951 g/d; $\sigma_Y = 408.9$), whereas the mean model predicted DMI was 958.1 g/d (range of 474 to 1,429 g/d; $\sigma_X = 306.5$). To estimate σ^2_ε, a subset of daily DMI on 7 animals were used. A random regression model with repeated measurements (first-order autoregressive) was fitted, yielding a residual error variance of 21,655. This value served as the estimate of daily DMI variance within an animal. Because the observations on the 67 kids were the average DMI over 76 d, an estimate of the precision of the observations $\left(\sigma^2_\varepsilon\right)$ was calculated as 21,655 ÷ 76 = 284.9. The precision of the NRC (2007) computer model was estimated using the published standard errors of the parameter estimates in [41] and estimated SE of BW and ME in the studies through a Monte Carlo method, assuming a multivariate normal distribution of all parameters (Fan et al., 2002). Thus, for this example, we have

$$\begin{aligned} \sigma^2_\eta &= 166{,}944 & & & S_{XX} &= 6{,}202{,}101 \\ \sigma^2_\varepsilon &= 284.9 & & \text{and} & S_{YY} &= 11{,}036{,}463 \\ \sigma^2_\xi &= 93{,}863 & & & S_{XY} &= 7{,}286{,}558 \\ \sigma^2_\delta &= 79.14 & & & \lambda &= 3.6 \end{aligned}$$

Using [23], the information size is calculated as

$$\kappa^2 = \frac{166{,}914}{284.9} + \frac{93{,}863}{79.14} = 1{,}771.9.$$

From Table 1, this κ^2 should provide a 95% CI width ω_β somewhere around 0.15 with a 95% probability. That is, we should be able to conclude practical equivalence between the model and the observations for approximately $0.85 < \beta < 1.176$.

Results are reported in Table 4. Details on the equations used and calculations involved follows.

Parameter Estimates. The estimate of β is found using [7]

$$\hat{\beta} = \frac{11{,}036{,}463 - (3.6 \times 6{,}202{,}101) + \sqrt{\left[11{,}036{,}463 - \left(3.6 \times 6{,}202{,}101\right)\right]^2 + \left(4 \times 3.6 \times 7{,}286{,}558^2\right)}}{2 \times 7{,}286{,}558};$$

$$\hat{\beta} = 1.275.$$

The estimate of θ_S is calculated using [17]:

$$\hat{\theta}_S = \arctan\left(\frac{1.275}{\sqrt{3.6}}\right) = 0.5916.$$

Table 4. Statistics for the 2 examples: (1) NRC (2007) prediction model of DMI of growing dairy goats, and (2) NRC (2001) prediction model of microbial N flow to the duodenum of dairy cows

Statistics	Example 1: Goat DMI	Example 2: Microbial protein
Information size (κ^2)[1]	1,771	27.4
$\hat{\theta}_S$[2]	0.5916	0.9557
95% CI on θ_S	$0.5298 < \theta_S < 0.6534$	$0.8598 < \theta_S < 1.0516$
$\hat{\theta}$[3]	0.9056	1.067
95% CI on θ	$0.8379 < \theta < 0.9679$	$0.9790 < \theta < 1.151$
$\hat{\beta}$[4]	1.275	1.814
95% CI on β	$1.111 < \beta < 1.453$	$1.488 < \beta < 2.423$
$\hat{\alpha}$[5]	−425.8	203.8
Mean bias ($\overline{B}$, g/d)[6]	162.3	1.44
95% CI on $\overline{B}$ (g/d)	$113.3 < \overline{B} < 211.4$	$-8.40 < \overline{B} < 11.28$
Rejection criterion on mean bias[7]	$\lvert 162.3\rvert > 50.97$	$\lvert 144\rvert < 4.0$
Rejection criterion on linear bias[8]	$0.1066 > 0.0242$	$0.1753 > 0.0087$

[1]The information size is calculated using equation [23]. It is the sum of 2 ratios of the variance of the true values for the observations and the computer model divided by their respective precision. A large κ^2 indicates a large spread of observations and model predictions relative to their respective precision parameters.

[2]Where $\hat{\theta}_S$ is the estimate of the slope when the regression is expressed in polar form and the model predictions are transformed according to equation [16] when the precision ratio $\lambda \neq 1$. It is calculated using equation [17].

[3]Where $\hat{\theta}$ is the estimate of the slope when the regression is expressed in polar form but without transformation of the model predictions. It is calculated using equation [12].

[4]Where $\hat{\beta}$ is the estimate of the slope when the regression is expressed in the common slope-intercept format. It is calculated using equation [7].

[5]Where $\hat{\alpha}$ is the estimate of the intercept when the regression is expressed in the common slope-intercept format. It is calculated using equation [9].

[6]The mean bias is calculated using equation [10].

[7]The rejection criterion on the mean bias is calculated using equation [31].

[8]The rejection criterion on the linear bias is calculated using equation [28].

The parameter α is estimated using [9]:

$$\hat{\alpha} = 795.8 - (1.275 \times 958.1) = -425.8.$$

Last, the mean bias $\overline{B}$ is estimated using [10]:

$$\hat{\overline{B}} = 958.1 - 795.8 = 162.3.$$

Equivalence Tests. First, we need to define ψ_S. Suppose that we would be happy with $0.8 < \beta < \beta_U$. From Table 2, $\psi_S \approx 0.092009$ for $\beta_L = 0.8$ and $\lambda = 3.5$, but we can also calculate the exact ψ_S using [29]:

$$\psi_S = \arctan\left(\frac{\sqrt{3.6} - 0.8\sqrt{3.6}}{0.8 + 3.6}\right) = 0.0860,$$

which yields an upper bound for β (equation [30]):

$$\beta_U = \frac{\sqrt{1.642} + 1.642\tan(0.0860)}{\sqrt{1.642} - \tan(0.0860)} = 1.219.$$

Using [19], we can also calculate $\phi_{\gamma/2}(\lambda)$:

$$\phi_{\gamma/2}(\lambda) = \frac{1}{2}\arcsin\left(1.997 \times \frac{2}{\sqrt{67-2}} + \sqrt{\frac{3.6 \times \left(11,036,463 \times 6,202,101 - 7,286,558^2\right)}{\left(11,036,463 - 3.6 \times 6,202,101\right)^2 + \left(4 \times 3.6 \times 7,286,558^2\right)}}\right);$$

$$\phi_{\gamma/2}(\lambda) = 0.0618.$$

The equivalence test on θ_S is given by [28]:

$$\left|0.5916 - 0.4850\right| = 0.1066 > (0.0860 - 0.0618) = 0.0242.$$

Hence, we cannot reject the null hypothesis and must conclude that the computer model and the observations are not average equivalent.

For the overall bias assessment, we must determine a range of tolerance (i.e., a reasonable value for $\psi_{\bar{B}}$ in [31]). Suppose that we would be satisfied if the overall bias was less than 2 × SEM ≈ 100 g/d. So, we set $\psi_{\bar{B}} = 100$. From [21],

$$SE_{\bar{B}} = \sqrt{\frac{6,202,101 + 11,036,463 - (2 \times 7,286,558)}{67 \times (67-1)}} = 24.55.$$

From [32],

$$\phi_{\gamma/2,\bar{B}} = 1.997 \times 24.55 = 49.03.$$

And last, using [31],

$$\left|162.3\right| > 100 - 49.03 = 50.97.$$

Hence, we cannot reject the null hypothesis and we must conclude a significant overall bias between the computer model predictions and the observations.

Confidence Intervals. Using [19], we first calculate $\phi_{\gamma/2}(\lambda)$:

$$\phi_{\gamma/2}(\lambda) = \frac{1}{2}\arcsin\left(1.997 \times \frac{2}{\sqrt{67-2}} + \sqrt{\frac{3.6 \times \left(11,036,463 \times 6,202,101 - 7,286,558^2\right)}{\left(11,036,463 - 3.6 \times 6,202,101\right)^2 + \left(4 \times 3.6 \times 7,286,558^2\right)}}\right);$$

$$\phi_{\gamma/2}(\lambda) = 0.0618.$$

Then using [18], we get the 95% CI for θ_S:

$$0.5916 - 0.0618 < \theta_S < 0.5916 + 0.0618$$
$$0.5298 < \theta_S < 0.6534$$

The 95% CI for β is then calculated using [20]:

$$1.8974\tan(0.5298) < \beta < 1.8974\tan(0.6534)$$
$$1.111 < \beta < 1.453$$

Last, the 95% CI on $\bar{B}$ is calculated using [22]:

$$162.4 - 49.03 < \bar{B} < 162.4 + 49.03$$
$$113.3 < \bar{B} < 211.4$$

Graphical Presentation of Results. Data and estimated regression lines are presented graphically in Figure 4. The GePReM regression is, as expected, within the area bounded by the OLS and IR lines, and, also as expected, closer to OLS than IR in this instance because $\lambda > 1$. The line of perfect equivalence is also shown. In this example, we conclude that the mean bias between the computer model and the observation (+162.4 g/d) is

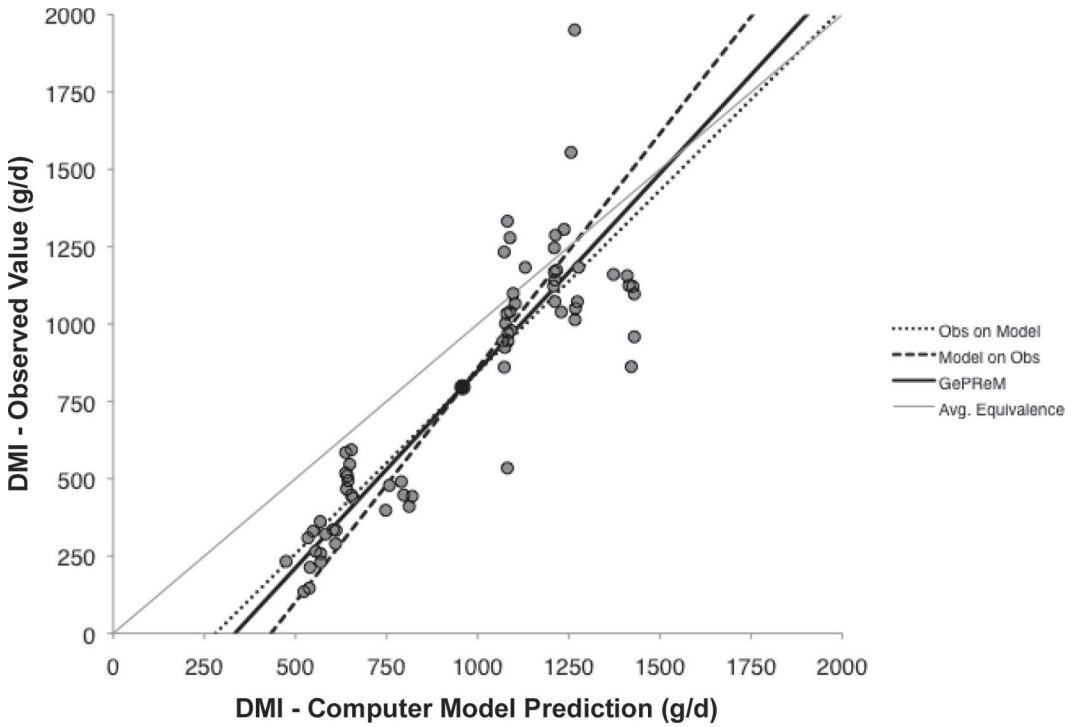

Figure 4. Observed DMI regressed on the NRC (2007) computer model predicted DMI for the growing goats example. Obs on Model is the regression line obtained by ordinary least-squares regression; Model on Obs is the regression line obtained by inverse regression; GePReM is the regression line obtained using the generalized projection regression method detailed in this paper; Avg. equivalence is the simple line of unity.

significantly greater than the boundary of practical equivalence (±100 g/d). The linear bias also significantly deviates from practical equivalence. The estimated slope being >1.0 with a negative $\hat{\alpha}$ indicates that the differences between the computer model predictions and the observations become smaller as computer model predictions increase. This difference becomes null when the computer model prediction equals 1,548 g/d.

Example 2: Prediction of Microbial N Flow to the Duodenum in Dairy Cattle

The data for the second example were digitally extracted by St-Pierre (2003) from Figure 5-6 of NRC (2001), as was done in a prior publication (St-Pierre, 2003). The complete data set is provided in Supplemental Table S2 (http://dx.doi.org/10.3168/jds.2015-10032). In total, 256 observations of the flow of microbial N to the duodenum of dairy cattle were gathered from 99 peer-reviewed publications. Observed N flows to the duodenum averaged 243.1 g/d (range of 85.5 to 442.2 g/d; $\sigma_Y = 74.89$ g/d). The model to be assessed was that of NRC (2001), which predicts N flow to the duodenum based on the calculated discounted TDN of the diet. The computer model predictions averaged 246.4 g/d (range from 119.3 to 351.9 g/d; $\sigma_X = 50.36$ g/d). The precision parameter for the observations $\left(\sigma_\varepsilon^2\right)$ was estimated using the median of the SEM reported across publications. The precision parameter for the computer model was estimated by a Monte Carlo method using the sample diet of a 680-kg Holstein cow producing 35 kg/d of milk at 3.5% fat, as found in Table 14-7 of NRC (2001), and the compositional variances reported in Table 15-1 of NRC (2001). In this example, we have the following values to work with:

$$\begin{aligned} \sigma_\eta^2 &= 5,364.5 & & & S_{XX} &= 641,635 \\ \sigma_\varepsilon^2 &= 148.19 & & \text{and} & S_{YY} &= 1,521,560 \\ \sigma_\xi^2 &= 2,445.6 & & & S_{XY} &= 515,152 \\ \sigma_\delta^2 &= 90.25 & & & \lambda &= 1.64 \end{aligned}$$

We note in passing that a value of $\lambda = 1.64$ indicates that the precision of the computer model predictions is close to that of the measurements. A value of $\lambda = 1$ would indicate identical precision.

We start with the calculation of the information size, using [23]:

$$\kappa^2 = \frac{2,445.6}{90.25} + \frac{5,364.5}{148.19} = 27.1 + 36.2 = 63.3.$$

From Table 1, the information size from these data should be sufficient for estimating the slope β within a width of 0.20. Therefore, we will select a low bound $\beta_L = 0.8$ to conduct the equivalence tests.

Results are reported in Table 4. Details on the equations used and the calculations involved are as follows.

Parameter Estimates. We use equation [7] to find the estimate of β:

$$\hat{\beta} = \frac{1,521,560 - (1.642 \times 641,635) + \sqrt{\left[1,521,560 - \left(1.642 \times 641,635\right)\right]^2 + \left(4 \times 1.642 \times 515,152^2\right)}}{2 \times 515,152};$$

$$\hat{\beta} = 1.814.$$

Using [17], we then estimate θ_S:

$$\hat{\theta}_S = \arctan\left(\frac{1.814}{\sqrt{1.642}}\right) = 0.956.$$

The intercept α is estimated using [9]:

$$\hat{\alpha} = 243.2 - \left(1.814 \times 246.4\right) = -203.8.$$

Last, the estimate of the overall bias $\bar{B}$ is calculated using [10]:

$$\hat{\bar{B}} = 246.4 - 243.2 = 3.2.$$

Equivalence Tests. To conduct equivalence tests, we first need to calculate ψ_S using [29]:

$$\psi_S = \arctan\left(\frac{\sqrt{1.642} - \left(0.8 \times \sqrt{1.642}\right)}{0.8 + 1.642}\right) = 0.1046,$$

which yields an upper bound for practical equivalence on β (equation [30]):

$$\beta_U = \frac{\sqrt{1.642} + 1.642 \tan(0.0504)}{\sqrt{1.642} - \tan(0.0504)} = 1.236.$$

Using [19], we calculate $\phi_{\gamma/2}(\lambda)$:

$$\phi_{\gamma/2}(\lambda) = \frac{1}{2}\arcsin\left(1.969 \times \frac{2}{\sqrt{254}} \times \sqrt{\frac{1.642 \times \left(1,521,560 \times 641,635 - 515,152^2\right)}{\left[1,521,560 - \left(1.642 \times 641,635\right)\right]^2 + \left(4 \times 1.642 \times 515,152^2\right)}}\right);$$

$$\phi_{\gamma/2}(\lambda) = 0.0959.$$

The equivalence test on θ_S is conducted using [28]:

$$\left|0.9557 - 0.7804\right| = 0.1753 > 0.1046 - 0.0959 = 0.0087.$$

The null hypothesis cannot be rejected. We must conclude that the computer model and the observations are not average equivalent.

The observed microbial N flow to the duodenum averaged 243.2 g/d; we arbitrarily chose a value of $\psi_{\bar{B}} = 243.2 \times 0.05 = 12.2$ g/d (i.e., a tolerance on the overall bias equal to 5% of the observed values). From [21]:

$$SE_{\bar{B}} = \sqrt{\frac{641,635 + 1,521,560 - (2 \times 515,152)}{256(256-1)}} = 4.166.$$

From [32]:

$$\phi_{\gamma/2,\bar{B}} = 1.969 \times 4.166 = 8.20.$$

Therefore, using [31]:

$$1.44 < 12.2 - 8.2 = 4.0,$$

and we reject the null hypothesis that the overall bias is greater than |12.2|: the overall bias is deemed practically insignificant.

Confidence Intervals. We first calculate $\phi_{\gamma/2}(\lambda)$ using [19]:

$$\phi_{\gamma/2}(\lambda) = \frac{1}{2}\arcsin\left(1.969 \times \frac{2}{\sqrt{254}} \times \sqrt{\frac{1.642 \times \left(1,521,560 \times 641,635 - 515,152^2\right)}{\left[1,521,560 - \left(1.642 \times 641,635\right)\right]^2 + \left(4 \times 1.642 \times 515,152^2\right)}}\right),$$

$$\phi_{\gamma/2}(\lambda) = 0.0959.$$

The 95% CI on θ_S is calculated using [18]:

$$0.9557 - 0.0959 < \theta_S < 0.9557 + 0.0959$$

$$0.8598 < \theta_S < 1.052.$$

The 95% CI on β is then easily calculated using [20]:

$$1.281 \tan(0.8598) < \beta < 1.281 \tan(1.052)$$

$$1.488 < \beta < 2.423.$$

Finally, the 95% CI on $\bar{B}$ is calculated using [22]:

$$1.44 - 8.20 < \bar{B} < 1.44 + 8.20$$

$$-6.76 < \bar{B} < 9.64.$$

The computer model is a practical shift equivalent to the observations.

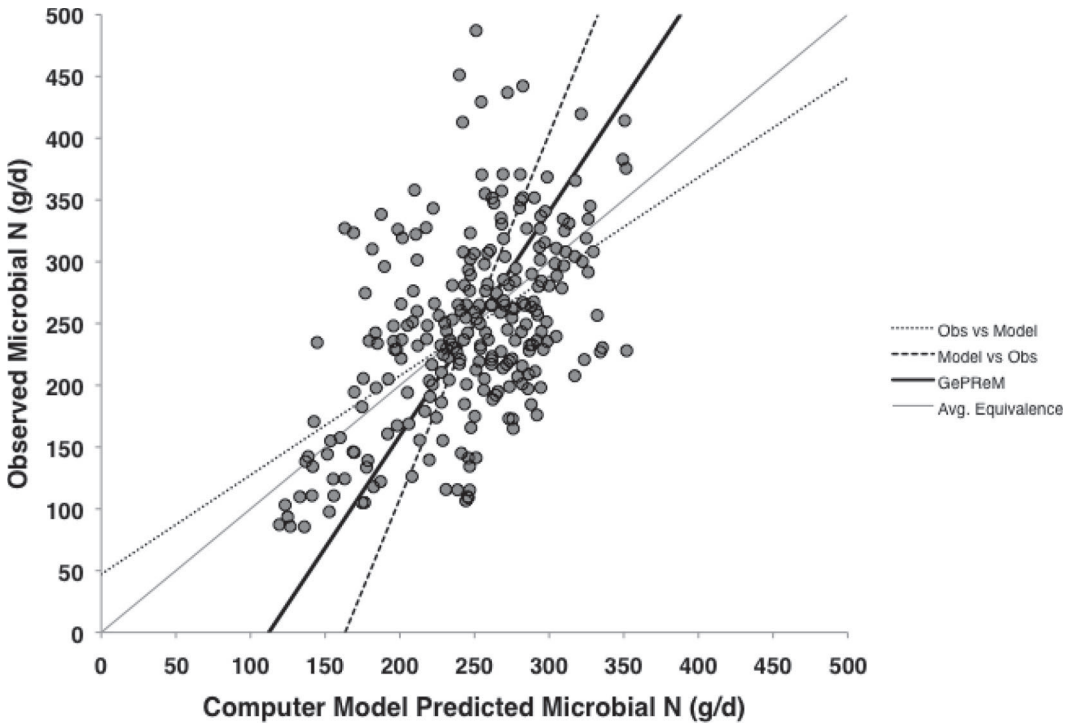

Figure 5. Observed microbial N flow to the duodenum regressed on the NRC (2001) computer model predicted value for the dairy cattle example. Obs on Model is the regression line obtained by ordinary least-squares regression; Model on Obs is the regression line obtained by inverse regression; GePReM is the regression line obtained using the generalized projection regression method detailed in this paper; Avg. equivalence is the simple line of unity.

Graphical Presentation of Results. Regression lines and data are shown graphically in Figure 5. As it should, the GePReM line is within the space spanned by the OLS and IR lines. This visual appraisal confirms the lack of an overall bias but the important linear bias by the computer model. This is in contrast to the conclusions reached by St-Pierre (2003), who used a residual regression approach, which lacks sensitivity and, more importantly, frames the hypotheses in a traditional difference test rather than an equivalence test. Hence, the computer model would be called practical shift equivalent to the observations.

Illustrating the GePReM Properties Using the Two Examples

In the first section of this paper, we argued that an ideal computer assessment method must exhibit 4 characteristics. Using the results from the 2 examples, we can illustrate how GePReM incorporates these. The first characteristic of an ideal method is that it accounts for the stochastic nature of observations and model predictions, which is explicit in GePReM as it requires estimates of the precision parameters for the observations $\left(\sigma_{\varepsilon}^2\right)$ and for the model predictions $\left(\sigma_{\delta}^2\right)$. In the first example (prediction of DMI in growing goats), $\sigma_{\varepsilon}^2 = 284.9$ and $\sigma_{\delta}^2 = 79.14$, whereas in the second example (prediction of microbial N flow to the duodenum in dairy cattle), $\sigma_{\varepsilon}^2 = 148.19$ and $\sigma_{\delta}^2 = 90.25$. Thus, the stochastic nature of both the measurements and the model predictions is explicit in GePReM.

The second property of an ideal method is that it sets correct null hypotheses. As opposed to the conventional null hypothesis that would set, as the defaults, that the computer model has no mean bias and no linear bias, the null hypotheses set by GePReM were that the model mean bias and linear bias exceed predetermined tolerances at a probability $\gamma/2 = 0.05$. In the first example, we did set a tolerable mean bias of 100 g/d; the actual bias was 162.3 g/d. Hence, we could not reject the null hypothesis and concluded that the model predictions have a mean bias. As for the linear bias, our tolerance on β was set at $0.85 < \beta < 1.176$. The calculated slope was $\beta = 1.275$. Hence, we could not reject the null hypothesis and had to conclude a significant linear bias by the model predictions (i.e., computer model predictions and observations are not average equivalent). In the second example, the tolerable mean bias was set at 12.2 g/d. The calculated mean bias was 3.3 g/d. We rejected the null hypothesis

that the bias was greater than |12.2| and concluded that the overall bias was practically insignificant. The tolerance on the linear bias was set at $0.80 < \beta < 1.235$. The actual slope was $\beta = 1.814$. Hence, we could not reject the null hypothesis and had to conclude that the model predictions showed a significant linear bias: the computer model predictions and the observations are not average equivalent.

The third property of an ideal assessment method is that the same conclusions are reached regardless of whether computer predictions or the observations are considered as X or Y. This is an explicit and fundamental property of GePReM: the same conclusions are reached in the 2 examples if we interchange observations and model predictions in all the equations. Concluding that the model predictions are not average equivalents to the observations is the same as concluding that the observations are not average equivalent to the computer predictions: they are simply not measuring the same thing. If we have more confidence in the measurements, which is generally the case, then we would simply conclude that the model predictions cannot replace (i.e., they are not average equivalent) the observations in practice.

The last property of an ideal assessment method is that it provides interpretable statistics on precision and accuracy: GePReM forces scientists to determine the precision of both measurements and model predictions. Recall that in the first example, the precision parameters were $\sigma^2_{\varepsilon} = 284.9$ for the measurements and $\sigma^2_{\delta} = 79.14$ for the model predictions, yielding a precision ratio $\lambda = 3.6$. In the second example, $\sigma^2_{\varepsilon} = 148.19$ and $\sigma^2_{\delta} = 90.25$, resulting in a precision ratio $\lambda = 1.64$. The interpretation of this ratio is straightforward: the model predictions are 3.6 times more precise than the measurements in example 1 and 1.64 times more precise in example 2. In both examples, the model is more precise than the measurements. Statistics on accuracy also have a straightforward interpretation in GePReM. The mean bias is expressed in the same units as those of the measurements and model predictions. In the first example, the mean bias $\hat{\bar{B}}$ was estimated at 162.3 g/d, a value equal to 20% of the average observed DMI. In the second example, the mean bias was estimated at 3.2 g/d, which equates to 1.3% of the average measurement. The linear bias is expressed as a slope; $\beta = 1.0$ indicates an absence of linear bias. In the first example, the estimated slope $\hat{\beta} = 1.275$, a value that is significantly outside the tolerance region that was set at $0.85 < \beta < 1.176$. In the second example, $\hat{\beta} = 1.814$, which clearly lies outside the tolerance region, which was set at $0.80 < \beta < 1.235$. The conclusion from both examples is that even if we were to adjust the models for the mean biases, model predictions and observations are not on the same scale—they are not measuring the same thing.

FUTURE DEVELOPMENTS

Additional work is needed to extend the GePReM to more complex situations. First, observation data are frequently gathered across many studies. Because these data are inherently imbalanced, it might be important that the random effect of "study" be incorporated in the assessment model, resulting in a mixed-effect model (St-Pierre, 2001). Maximum likelihood estimates of these random effects in combination with generalized projection regression for the fixed effects are currently unknown. One possibility would be to use a quasi-REML approach, where the fixed effects would be absorbed based on generalized projections and the resulting modified observations used for ML estimation of the random effects. This process would then be iterated until stable estimates of both the fixed and random effects are obtained. On the other hand, the realized values (i.e., BLUP) of future "studies" (i.e., when using the model in practice) are not known. Hence, ignoring their effect in a GePReM approach may arguably be correct if our interest is in assessing the model for its ability to predict future values. This issue is in need of additional research.

Second, when means of observations rather than individual observations are used, the precision of the observations and computer model predictions can depend on the size and design of the experiment that generated the observed means. This could have been an issue in our second example, but because all observations were means from Latin square designs, this was probably of trivial importance. Assigning different weights to the observations could partially alleviate the unequal precision, but the effectiveness of this remedy is unknown at this time.

Third, GePReM handles only overall and linear comparisons between model predictions and observations. Much additional information could be gathered through residual analysis. Residuals calculated according to equation [35] could be subjected to statistical process control charts to detect nonlinear or unusual patterns.

Finally, observations can be gathered from multiple measurements on the experimental units, leading to correlated errors. The GePReM does not account for the correlated errors and the consequences of this are currently unknown.

ACKNOWLEDGMENTS

Salary and research support was provided by state and federal funds appropriated to the Ohio Agricultural Research and Development Center, The Ohio State University (Manuscript 50/14AS). The author thanks Izabella Teixeira (Universidade Estadual Paulista, São Paulo, Brazil) for sharing her research data and providing a welcoming and quiet work refuge and words of encouragement during the writing of this manuscript. Many thanks to my colleagues Izabella Teixeira and William Weiss from The Ohio State University, Wooster campus, and Joanne Knapp of Fox Hollow Consulting LLC (Columbus, OH) for their comments on a prior version of this manuscript.

REFERENCES

Altman, D. G., and J. M. Bland. 1983. Measurements in medicine: The analysis of methods comparison studies. Statistician 32:307–317.

Anderson, T. W. 1976. Estimation of linear functional relationships: Approximate distribution and connection with simultaneous equations in econometrics. J. Royal Stat. Soc. B 38:1–36.

Anderson, T. W. 1984. Estimating linear statistical relationships. Ann. Stats 12:1–45.

Berger, R. L., and J. C. Hsu. 1996. Bioequivalence trials, intersection union tests and equivalence confidence sets. Stat. Sci. 11:283–319.

Bibby, J., and H. Toutenburg. 1977. Prediction and Improved Estimation in Linear Models. Wiley, Berlin, Germany.

Carroll, R. J., and D. Ruppert. 1996. The use and mis-use of orthogonal regression in linear errors-in-variables models. Am. Stat. 50:1–6.

Casella, G., and R. L. Berger. 1990. Statistical Inference. Wadsworth, Pacific Grove, CA.

Creasy, M. A. 1956. Confidence limits for the gradient in the linear functional relationship. J. Royal Stat. Soc. B 18:65–69.

Draper, N. R., and H. Smith. 1988. Applied Regression Analysis. 3rd ed. John Wiley and Sons Inc., New York, NY.

Fan, X., A. Felsovalyi, S. A. Sivo, and S. C. Keenan. 2002. SAS for Monte Carlo Studies: A Guide for Quantitative Researchers. SAS Institute Inc., Cary, NC.

Fuller, W. A. 1987. Measurement Error Models. John Wiley & Sons Inc., New York, NY.

Gleser, L. J., and J. T. Huang. 1987. The nonexistence of a 100(1-α)% confidence set of finite expected diameter in errors-in-variable and related models. Ann. Stat. 15:1351–1362.

Hozo, S. P., B. Djulbegovoc, and I. Hozo. 2005. Estimating the mean and variance from the median, range and the size of a sample. BMC Med. Res. Methodol. 5:13–25.

Kendall, M., and A. Stuart. 1979. The Advanced Theory of Statistics, Vol II: Inference and Relationships, 4th ed. Macmillan, New York, NY.

Lin, L. I.-K. 1989. A concordance correlation coefficient to evaluate reproducibility. Biometrics 45:255–268.

Lunn, D., D. Spiegelhalter, A. Thomas, and N. Best. 2009. The BUGS project: Evolution, critique and future directions. Stat. Med. 28:3049–3067.

Madansky, A. 1959. The fitting of straight lines when both lines are subject to error. J. Am. Stat. Assoc. 54:173–205.

Mandel, J. 1978. Accuracy and precision evaluation and interpretation of analytical results. Pages 243–298 in Treatise on Analytical Chemistry, Part I, Theory and Practice, Volume I. I. Kolthoff and P. J. Elving, ed. Wiley, New York.

Marino, S., I. A. Hogue, C. J. Ray, and D. E. Kirschner. 2008. A methodology for performing global uncertainty and sensitivity analysis in systems biology. J. Theor. Biol. 254:178–196.

Mayer, D. G., M. A. Stuart, and A. J. Swain. 1994. Regression of real-world data on model output: An appropriate overall test of validity. Agric. Syst. 45:93–104.

Mitchell, P. L. 1997. Misuse of regression for empirical validation of models. Agric. Syst. 54:313–326.

National Research Council. 2001. Nutrient Requirements of Dairy Cattle. 7th rev. ed. Natl. Acad. Sci., Washington, DC.

National Research Council. 2007. Nutrient Requirements of Small Ruminants. Natl. Acad. Sci., Washington, DC.

Solari, M. E. 1969. The maximum likelihood solution of the problem of estimating a linear functional relationship. J. Royal Stat. Soc. B 31:372–375.

St-Pierre, N. R. 2001. Invited Review: Integrating quantitative findings from multiple studies using mixed model methodology. J. Dairy Sci. 84:741–755.

St-Pierre, N. R. 2003. Reassessment of biases in predicted nitrogen flows to the duodenum by NRC 2001. J. Dairy Sci. 86:344–350.

St-Pierre, N. R. 2015a. Statistical issues in nutritional modeling. Pages 111–125 in Modelling in Pig and Poultry Production. N. Sakomura and R. Gous, ed. CABI International, London, UK.

St-Pierre, N. R., and C. S. Thraen. 1999. Invited. Animal grouping strategies, sources of variation, and economic factors affecting nutrient balance on dairy farms. J. Dairy Sci. 82(Suppl. 2):72–83.

Tan, C. Y., and B. Iglewicz. 1999. Measurement-methods comparison and linear statistical relationship. Technometrics 41:192–201.

Tedeschi, L. O. 2006. Assessment of the adequacy of mathematical models. Agric. Syst. 89:225–247.

Teixeira, I. A. M. A., N. St-Pierre, K. T. de Resende, and A. Cannas. 2011. Prediction of intake and average daily gain by different feeding systems for goats. Small Rumin. Res. 98:93–97.

Theil, H. 1961. Economic forecasts and policy. Pages 6–48 in Contributions to Economic Analysis, 2nd ed. R. Strotz, J. Tinbergen, P. J. Verdoorn, and H. J. Witteveen, ed., North Holland, Amsterdam, Holland.

Thornley, J. H. M., and J. France. 2007. Mathematical Models in Agriculture: Quantitative Methods for the Plant, Animal and Ecological Sciences. CABI, Wallingford, England.

Van Belle, G. 2002. Statistical Rules of Thumb. John Wiley & Sons Inc., New York, NY.

Warton, D. I., I. J. Wright, D. S. Falster, and M. Westoby. 2006. Bivariate line-fitting methods for allometry. Biol. Rev. Camb. Philos. Soc. 81:259–291.

Willassen, Y. 1979. Two clarifications on the likelihood surface in functional models. J. Mult. Anal. 9:138–149.